Eximia journal

www.eximiajournal.ro

Vol. 14/2025

PLUS COMMUNICATION

International Communication & PR

Eximia Journal
Vol. 14, 97-103, January, 2025
ISSN: 2784-0735
www.eximiajournal.com

Use of The Principal Components in Case of The Problem multicollinearity (An Applied Study on the investment data in Sudan)

Ashraf Hassan Idris Brama[1], Mohammed Abdalwahab Mohammed Salim[2], Monastir Abbas Ahmed Mohammed[3], Alaa Alfadel Ahmed Abuzaid[4]

Department of Management Information System, Collage of Business & Economics, Qassim University, Buraydah, Kingdom of Saudi Arabia, Saudi Arabia

a.brama@qu.edu.sa, m.salim@qu.edu.sa, m.mahmmeed@qu.edu.sa , a.abuzid@qu.edu.sa

Abstract. This study aims to apply the model Principal component Analysis to overcome the problem of multicollinearity which appears in many estimated models. The core of this method is replacing the independence variables that suffer from the multicollinearity problem with principal components to solve this problem. The problem of multicollinearity arises in a regression model when the independent variables are highly correlated to each other. Multicollinearity does not limit the possibility of obtaining a good model or affect inferences about expected responses the hypotheses the principal components regression method has the capability of resolving the multicollinearity that exists between explained variables , It is possible to conclude that throughout our paper; the problem of multi-collinearity has been solved by using a principal component technique in the case of the panel regression model, by using investment data in the Sudan.

Keywords. Multicollinearity, principal components, regression, estimation, VIF

1. Introduction:

In modelling the economic growth, we are often constrained by models that do not meet the assumptions, one of which is multicollinearity. This occurs because the data obtained is taken from uncontrollable circumstances. These cases can cause difficulty in separating the influence of each independent variable (X) on the response variable (Y), so, we need a method to solve it. One method that can be used is Principal Component Regression (PCR). principal component analysis is the oldest and best-known technique of multivariate data analysis. It was first coined by Pearson (1901) and developed independently by Hotelling (1933). Like many other multivariate methods, it was not widely accepted nor used until the advent of electronic computers, but it is now well entrenched in virtually every statistical software packages.

2. Research problem:

The ordinary least squares method gives the best unbiased linear estimation with the least variance of the model parameters. One of the issues

Eximia Journal
Vol. 14, 97-103, January, 2025
ISSN: 2784-0735
www.eximiajournal.com

One of the issues that may arise when using this method is the absence of one of the assumptions of the linear model, which requires that there is no full or partial linear correlation between two or more explanatory variables, which leads to the emergence of an issue called the linear interference issue, which causes poor parameter estimates with inflated variances. Thus, the results of hypothesis tests are unreliable, as well as finding a function that is a criterion for total investment in Sudan.

3. Research significant:

The importance of the research lies in addressing the problem of multicollinearity among explanatory variables and finding appropriate solutions and treatments for it. One of the most important methods for addressing this issue is the principal component regression.

4. Research hypotheses:

a) The general model of the data gives better results than the results obtained by other estimation methods, the explanatory variables are independent (there is no collinearity between them)

b) The principal components regression method has the capability of resolving the multicollinearity that exists between explained variables

5. Principal Component Analysis (PCA):

PCA is the most widely used multivariate method.

PCA uses the correlation structure of original variables and derives p linear combinations which are uncorrelated. Each PC provides unique information about the data. Although 'p' principal components are derived, the first 'k' principal components are expected to explain most of the variability in the data. From p original variables: X_1, X_2, …….,X_n derive p new variables $b_1, b_2, \ldots., b_n$ □

$y_1 = a_{11}x_1 + a_{12}x_2 + a_{13}x_3 + \cdots + a_{1p}x_p$

$y_2 = a_{21}x_1 + a_{22}x_2 + a_{23}x_3 + \cdots + a_{2p}x_p$

$y_3 = a_{31}x_1 + a_{32}x_2 + a_{32}x_3 + \cdots + a_{3p}x_p$

.

.

$y_p = a_{p1}x_1 + a_{p2}x_2 + a_{p3}x_3 + \cdots + a_{pp}x_p$

such that Y_k are uncorrelated (orthogonal).

Variance-Covariance matrix can be used if variables are measured on the same units. Although, Correlation matrix is a better choice for PCA than the variance-covariance matrix. The variance-covariance matrix of the standardized variables is same as the correlation matrix of the original variables

6. Principal Component Regression (PCR):

PCR is a regression technique that combines elements of both PCA and linear regression. In PCR, PCA is first performed on the predictor variables to obtain a reduced set of principal components. These principal components are then used as predictors in a linear regression model to predict the response variable. In the Principal Component Regression, the first k principal components are used as the independent variables instead of the original X(explanatory) variables. Each PC is a linear combination of all X variables. The final model is expressed in terms of original independent variables for ease of interpretation using back transformation. In the first step, the original p variables are transformed into a new set of orthogonal or uncorrelated variables called “Principal Components”. In the second step, after the elimination of the least important principal components, a multiple regression analysis of the response variable against the reduced set of principal components is performed using the

Eximia Journal
Vol. 14, 97-103, January, 2025
ISSN: 2784-0735
www.eximiajournal.com

OLS (Ordinary Least Square) estimation. In the third step, the model equation is back transformed in terms of the original variables.

The statistical model for Principal Component Regression (PCR) is as below.

$y = a_0 + a_1 pc_1 + a_2 pc_2 + \cdots + a_k pc_k + \acute{e}$

The Multiple Linear Regression technique is a predictive Modelling technique that includes one numeric dependent variable and several explanatory variables. Using the historical data, a mathematical equation is built that is used for predicting the outcome.

$Y = b_0 + b_1 x_1 + b_2 x_2 + \cdots + b_n x_n$

Where, Y is a dependent variable and X_1, X_2, ……,X_n are independent variables. $b_1, b_2, \ldots, b_n$ are the parameters of the Model and 'e' is the Random Error Component?

7. **Indications:**

There are some indicators of the presence of multicollinearity.

a) High pairwise correlations among independent variables. (Sufficient but not necessary)
b) Significant F value (based on ANOVA) but very few significant t values.
c) One of the eigenvalues of the correlation matrix of independent variables is close to zero.

8. **Variance Inflation Factor** (VIF):

Regression analysis multicollinearity is measured statistically using the Variance Inflation Factor (VIF). It measures the extent to which multicollinearity among the predictor variables increases the variance of the predicted regression coefficients. According to Fox.

Frequently, Variance Inflation Factor (VIF) is used to detect the presence of multicollinearity. VIF is calculated for each independent variable.

$VIF = \frac{1}{1- R^2}$

Where, R^2, is the coefficient of determination obtained by building a Regression Model using each independent variable as independent variable and the rest as the independent variables. Each R^2 measures how much variation in the independent variables is being explained by the other independent variables. If the value of R^2 is high, then there is a problem of multicollinearity. The problem is severe if any of the above R^2 is greater than or equal to 0.80 or VIF > 5.

$R^2 = \frac{\sum_{i=1}^{n}(\hat{y}_i - \bar{y})^2}{\sum_{i=1}^{n}(y_i - \bar{y})^2}$ $0 \leq R^2 \leq 1$

9. Data analysis:

Multiple regression model coefficients and variance inflation factor (VIF):

Table (1) illustrate the regression coefficients test and the inflation factor between the explanatory variables.

model	Parameters		VIF
	β	St. error	
constant	-5.65	8.987	
1	1.285	3.955	120.869
2	-2.85	1.109	21.735
3	-0.209	2.509	61.369
4	0.182	0.575	4.014
5	0.389	0.561	6.669
6	0.044	0.236	3.844
7	-0.864	1.143	2.559
8	2.095	3.753	88.100

Eximia Journal
Vol. 14, 97-103, January, 2025
ISSN: 2784-0735
www.eximiajournal.com

9	0.067	0.769	4.371

IBM SPSS 25

from above table that most of the independent variables suffer from multicollinearity because most of the independent variables have variance inflation factor value greater than 10.

10. Test of significant model:

$$H_0: \beta_1 = \beta_2 = \cdots = \beta_9$$
$$H_1: \beta_1 \neq \beta_2 \neq \cdots \neq \beta_9$$

Table (2) illustrate the ANOVA for the significant model

S. O. V	SS	DF	MS	F	Sig.
Regressions	7.131	9	0.792	3.520	**0.045**
Error	1.801	8	0.225		
Total	8.932	17			

IBM SPSS 25

from above table the significant value (0.045) less than the probability value (0.05) therefore the model multilinear regression its statistically significant.

11. Test of residual normality

H_0: Residual with normal distribution

H_1: Residual with not normal distribution

Table (3) illustrate Test of the Residual Normality

Test	Kolmogorov-Smirnov			Shapiro-Wilk		
Residual	Statistic	Df	Sig.	Statistic	Df	Sig.
	0.141	18	0.200	0.967	18	0.742

IBM SPSS 25

from above table we note all significant value its grate than the probability value (0.05) therefore the residual values, with the normal distribution.

Table (4) illustrate the Correlation coefficient

R	R^2	Adjusted R^2
0.894	0.798	**0.572**

IBM SPSS 25

from above table we note the correlation coefficient (0.894) therefore its high correlation between variables.

12. Test of the fine multicollinearity with VIF:

Table (5) illustrate the principal component analysis:

Variable	1	2	3	8
1	----	0.911	0.960	0.990
2	0.911	------	0.942	0.853
3	0.960	0.942	--------	0.956
8	0.990	------	0.956	------

IBM SPSS 25

Table (6) illustrate the Test Kaiser and Bartlett's with Chi -Square:

Test	Statistic	Df	Sig.
Kaiser	0.745	-	-
Bartlett's	192.415	55	0.000

Eximia Journal
Vol. 14, 97-103, January, 2025
ISSN: 2784-0735
www.eximiajournal.com

IBM SPSS 25

Table (7) illustrate the Eigenvalue:

Eigenvalue	λ_1	λ_2	λ_3	λ_4	λ_5	λ_6	λ_7	λ_8	λ_9	Total
value	5.68	1.44	0.77	0.53	0.37	0.16	0.04	0.01	0.005	9

IBM SPSS 25

a) Eigenvalue (9) which is the number of explanatory variables
b) The eigenvalue ($\lambda_7 = 0.039, \lambda_8 = 0.0011, \lambda_9 = 0.005$) its vary low eigenvalue and these indicate multicollinearity between explanatory variables.

Coefficient matric (K) = $\sqrt{\frac{\lambda_1}{\lambda_9}} = \sqrt{\frac{5.68}{0.005}} = 33.7$

Where (K)> 30 there is indicate multicollinearity between explanatory variables.

Table (8) illustrate the principal component:

Variable	Components	
	Component1	Component2
Z_1	0.9560	0.60
z_2	0.9250	0.301
z_3	0.9730	0.151
z_4	-0.110	-0.835
z_5	0.794	0.086
z_6	0.6810	0.519
z_7	0.075	0.818
z_8	0.952	0.0130
z_9	0.686	0.3940

IBM SPSS 25

Table (9) illustrate the principal component matric:

component	1	2
1	0.942	0.335
2	0.335	-0.942

IBM SPSS 25

First Principal component:

Component1= $0.925z_2 + 0.973z_3 - 0.110z_4 + 0.794z_5 + 0681z_6 + 0.075z_7 + 0.952z_8 + 0.686z_9$

second Principal component:

Component2= $0.060z_1 + 0.301z_2 + 0.153z_3 - 0.835z_4 + 0.086z_5 + 0.519 + 0.818z_7 - 0.13z_8 + 0.394z_9$

13. Regression model of principal component:

Table (10) illustrate the Regression model of Principal component:

Model	Parameter		VIF
	β	S. E	
Constant	-0.139	0.221	-
Secondary indicators	-0.070	.080	3.342
Main indicators	0.001	0.001	3.342

IBM SPSS 25

14. Coefficient correlation of PCR model:

Eximia Journal
Vol. 14, 97-103, January, 2025
ISSN: 2784-0735
www.eximiajournal.com

Table (11) illustrate the correlation

R	R^2	Adjusted R^2
0.191	0.037	-0.041

IBM SPSS 25

From table (4) note decrease R value from (0.894) to (0.191) this indicates disappearance of multicollinearity among the explanatory variables

Result:

It is possible to conclude that throughout our paper; the problem of multi-collinearity has been solved by using a principal component technique in the case of the panel regression model, by using real data set was mentioned by

1. The regression model obtained through principal components is considered more efficient in the prediction process.
2. The principal component method is an effective approach to detecting the problem of multicollinearity among explanatory variables, as the principal components are always orthogonal.
3. The principal component method addressed the problem of multicollinearity among the explanatory variables because the principal components are orthogonal to each other. It also reduced the explanatory variables since the components are linear functions of the explanatory variables.
4. The regression model obtained using the Principal Component Regression (PCR) method is characterized by lower variance, which leads to accuracy in prediction results and can be relied upon for forecasting.
5. There are several tests and criteria used to detect the presence of multicollinearity among explanatory variables, and the principal component method is one of those criteria, which is considered easy to understand and apply.

Recommendations:

1. This method would be better used if we first know about the factor analysis.
2. Weakness in the factor analysis is the high subjectivity in determining factor. therefore, in determining the number of factors it is required an additional theory to form factors appropriately.
3. Using other methods for solving multicollinearity and comparing them to the methods under study, such as Bayesian Ridge Regression, Robust Ridge.
4. It is one of areas for our future work.

Conclusion:

The principal components method is an effective method for detecting and handling the problem of multicollinearity, it is easy to apply, and the regression model obtained by the principal component's method is more efficient in the prediction process.

References:

1. Fox, J. and Weisberg, S. (2018) An R Companion to Applied Regression. Sage Publications.
2. Imam, G., 2013. Multivariate Analysis Applications with IBM SPSS 21 Program Update Issue 7 PLS Regression. Diponegoro University, Semarang, Indonesia,.
3. Lazarsfeld, P. F. (1940). «Panel» Studies. Public Opinion Quarterly, 4 (1), 122. doi: https://doi.org/10.1086/265373

4. Andreß, H.-J. (2017). The need for and use of panel data. IZA World of Labor. doi: https://doi.org/10.15185/izawol.352
5. Baltagi, B. H. (2005). Econometric analysis of panel data. John Wiley & Sons Inc.Zulfikar, R. (2018). Estimation Model and Selection Method of Panel Data Regression: An Overview of Common Effect, Fixed Effect, and Random Effect Model. INA-Rxiv. doi: https://doi.org/10.31227/osf.io/9qe2b
6. Born, B., Breitung, J. (2014). Testing for Serial Correlation in Fixed-Effects Panel Data Models. Econometric Reviews, 35 (7), 1290–1316. doi: https://doi.org/10.1080/07474938.2014.976524
7. Greene, W. (2012). Econometric analysis. Prentice Hall.
8. Adeboye, N. O., Fagoyinbo, I. S., Olatayo, T. O. (2014). Estimation of the Effect of Multicollinearity on the Standard Error for Regression Coefficients. IOSR Journal of Mathematics, 10 (4), 16–20. doi: https://doi.org/10.9790/5728-10411620
9. Gujarati, D., Porter, C. (2008). Basic Econometrics. McGraw-Hill. Available at: https://cbpbu.ac.in/userfiles/file/2020/STUDY_MAT/ECO/1.pdf
10. Baltagi, B. H. (2021). Econometric analysis of panel data. Springer Cham, 424. doi: https://doi.org/10.1007/978-3-030-53953-5
11. Norliza Adnan, 1Maizah Hura Ahmad, Robiah Adnan.(2006).AComparative Study On Some Methods For Handling Multicollinearity
12. Mohd Sofiyan Sulaiman**,** Manal Mohsen Abood**,** Shanker Kumar Sinnakaudan, Mohd Rizal Shukor.(2019). Assessing and solving multicollinearity in sediment transport prediction models using principal component analysis
13. Multicollinearity in Regression Analysis: The Problem Revisited, https://doi.org/10.2307/1937887
14. Best subset selection for eliminating multicollinearity.(2016). Ryuhei Miyashiro, Kazuhide Nakata, Tomomi Matsui.
15. Implementation of PCA multicollinearity method to landslide susceptibility assessment: the study case of Kabylia region.(2023). Amel Kab, Lynda Djerbal & Ramdane Bahar.